二分タングルグラムの描き方 (列挙問題に対する計算の高速化と可視化)

AUTHOR(S):

岡本, 吉央

数理解析研究所講究録
第 1644 巻 2009 年 120-127

二分タングルグラムの描き方

Drawing Binary Tanglegrams: Recent Development

岡本 吉央 (Yoshio Okamoto) *

概 要

二分タングルグラム (binary tanglegram) とは根付き二分木の対で，その葉集合の間に 1 対 1 対応が存在し，対応する葉同士が辺によって結ばれているもののことである．例えば，系統学における応用では，それぞれの木を辺交差なく描き，木の間の辺の交差ができるだけ少なくなるように描くことが要求される．本稿では，二分タングルグラムの辺交差最小描画に対するアルゴリズムの最近の研究をまとめる．

1 序

タングルグラム (tanglegram) とは根付き木の対で，その葉集合の間に 1 対 1 対応があるもののことである [15]．根付き木の対を視覚的に比較したいという要求がソフトウェア工学，系統学，クラスタリングなどにはある．例えば，ソフトウェア工学では根付き木によってパッケージ=クラス=メソッド階層やプロジェクトの階層的分解を表現する．系統学では根付き木によって系統樹を表現し，異なる方法で得られた 2 つの系統樹を比較するためにタングルグラムが用いられる．クラスタリングにおいて，階層的クラスタリングの結果はデンドログラムと呼ばれる二分木のような構造で表現され，異なる手法によって得られた 2 つのデンドログラムの比較にタングルグラムが用いられる．系統樹の例を図 1 に示す．

本稿では断りがない限り，根付き木は二分木であると仮定し，葉の数は n であるとする．系統学やクラスタリングに現れる根付き木は二分木であるため，そのような応用を念頭に置いているわけである．2 つの二分木に対するタングルグラムを二分タングルグラムと呼ぶことにする．応用の視点からは，(a) 考えている木は平面上に交差なく描き，(b) 対応する葉同士を辺で結び，(c) 葉を結ぶ辺同士の交差を最小化することに意味がある．バイオインフォマティクスの文献 (例えば [15, 12]) に従って，これを**二分タングルグラム・レイアウト問題** (binary tanglegram layout problem, BTL) と呼ぶことにする[1]

*東京工業大学 (Tokyo Institute of Technology)

[1] Fernau, Kaufmann, and Poths [6] は「2 木交差最小化」(two-tree crossing minimization) と呼んでいる．

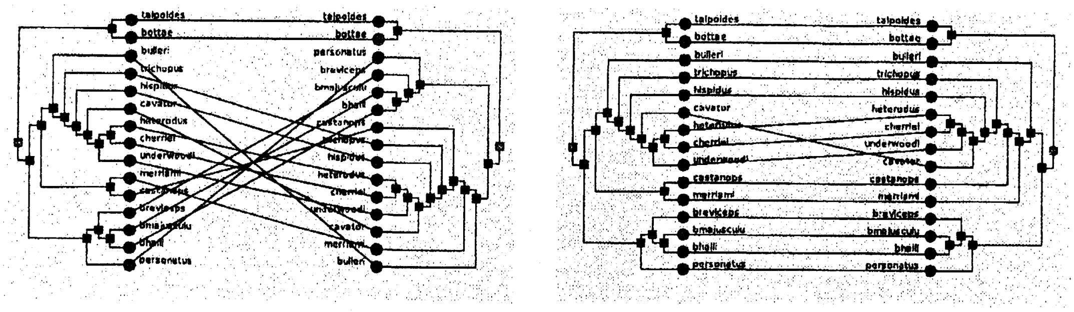

図 1: ホリネズミに対する 2 つの系統樹のタングルグラム [9]. (左) 交差の多い描画. (右) 交差の少ない描画.

関連研究 グラフ描画では，両面交差最小化問題 (two-sided crossing minimization problem, 2SCM) が階層的グラフ・レイアウトを計算する際に生じる重要な問題であると認識されている．そのようなレイアウトは Sugiyama, Tagawa, and Toda [16] によって導入され，階層的グラフの描画に対して広く用いられている．2SCM は二部グラフを描画の対象として，各部集合は平行な直線上に置かれなければならない．BTL と同様に，目的は交差数の最小化である．2SCM は NP 困難である [7]．片面交差最小化問題 (one-sided crossing minimization, 1SCM) は，片方の部集合の頂点の位置は既に固定されているバージョンである．1SCM は NP 困難である [5]．BTL と比較すると，1SCM や 2SCM では頂点の次数が高くてもよく，直線上の置き方も木の内部構造に制限されるわけではない．1SCM に対しては次が知られている．Eades and Wormald [5] によるメディアン・ヒューリスティクスは 3 近似であり，Nagamochi [13] の乱択アルゴリズムは 1.4664 近似である．Yamaguchi and Sugimoto [17] は $\gamma(\Delta)$ 近似を提案している．ただし，γ は 2 から 3 まで単調に増加する関数であり，$\gamma(4) = 2$ である．Dujmovič, Fernau, and Kaufmann [3] は $O^\star(1.4664^k)$ 時間の固定パラメータ・アルゴリズムを提案している．ここで，k は最適値である．$O^\star(\cdot)$ と言う記法は多項式因子を無視している．

形式的定義 木 T の葉集合を $L(T)$ で表す．葉の数が n の 2 つの根付き木 S と T が与えられたとする．そのとき，S と T の葉には唯一のラベルが付いているとする．すなわち，Λ をラベルの集合，例えば $\Lambda = \{1, \ldots, n\}$ として，全単射 $\lambda_S \colon L(S) \to \Lambda$ と $\lambda_T \colon L(T) \to \Lambda$ が存在するとする．このラベル付けから新たな辺集合 $\{uv \mid u \in L(S), v \in L(T), \lambda_S(u) = \lambda_T(v)\}$ が定義される．この集合の辺を**木間辺** (inter-tree edge) と呼ぶことにする．TL は S と T の無交差描画で，木間辺の間の交差を最小化するものを見つける問題である．ただし，辺はすべて直線分として描かれるものとし，$L(S)$ の葉は直線 $x = 0$ 上に置かれ，$L(T)$ の葉は直線 $x = 1$ 上に置かれるものとする．木 S と T はそれぞれ $x = 0$ の左側と $x = 1$ の右側に描かれるものとする．例が図 1 にある．葉に唯一のラベルが付いた 2 つの木 S と T が与えられたとき，そこから

得られる TL のインスタンスを $\langle S,T\rangle$ で表す．木 S と T が双方とも二分木である場合に制限した問題は BTL と呼ばれる．

本稿の構成 BTL に対して知られている結果を概観する．まず，BTL の片面バージョンについて議論する．その後，本来の BTL (すなわち両面バージョン) に議論を移し，交差なく描けるかどうかの判定について述べる．そして，交差最小化の困難性について議論し，近似アルゴリズム，固定パラメータ・アルゴリズムについて言及する．最後に実験に関する結果を見る．

2 片面バージョン

BTL の片面バージョンでは，片方の木は固定し，もう一方の木のレイアウトを変えることで交差最小化を図る．Dwyer and Schreiber [4] はこの問題に対する $O(n^2)$ 時間アルゴリズムを提案した．基本的なアイディアは動的計画法である．BTL のインスタンス $\langle S,T\rangle$ において，S の描画が固定されているとする．T の根の子を根とする 2 つの部分木を T_1, T_2 とすると，$\langle S,T\rangle$ に対する辺交差は次の 3 種類に分類できる．(a) T_1 の葉集合と T_1 の葉に対応する S の葉の集合を結ぶ木間辺の交差，(b) T_2 の葉集合と T_2 の葉に対応する S の葉の集合を結ぶ木間辺の交差，(c) T_1 の葉に接続する木間辺と T_2 の葉に接続する木間辺の間の交差．この中で，(a) と (b) の交差は再帰によって計算でき，(c) の交差は T_1 と T_2 の配置の仕方 (これは 2 通りしかない) のみによって定まる．したがって，再帰部分を動的計画法によって実装すると，解くべき部分問題の数は $O(n)$ となり，各部分問題は $O(n)$ 時間で解くことができるため，全体の計算量は $O(n^2)$ となる．

上で説明した Dwyer and Schreiber [4] のアルゴリズムでは各部分問題を解くとき (c) の交差を計算する部分で $O(n)$ 時間も費やしている．その部分を見直すことで，Fernau, Kaufmann, and Poths [6] は計算量を $O(n \log^2 n)$ に改善した．

3 交差なく描けることの判定

本節より BTL の両面バージョン (すなわち本来の BTL) を考える．始めに，交差なく 2 つの二分根付き木を描くことができるかどうか考える．

Fernau, Kaufmann, and Poths [6] は与えられた二分タングルグラムを辺交差なく描くことができるかどうか判定する問題が線形時間 $O(n)$ で解けることを以下のようにして示した．与えられたインスタンス $\langle S,T\rangle$ から次のような有向非閉路グラフ D を構成する．まず，D の頂点集合は S と T の頂点集合 (の合併) である．D の辺集合は S と T の辺集合と，S と T の木間辺の集合 (の合併) である．各辺の向きは次のように定める．まず，S の辺は

根から葉に向かって有向化する．S と T の木間辺は S から T に向かって有向化する．T の辺は葉から根に向かって有向化する．こうすることで，D の upward-planarity と $\langle S,T\rangle$ が木間辺の交差なく描けることが同値であることが分かる．ソースが 1 つしかない有向非閉路グラフの upward-planarity は線形時間で判定できるため [1]，タングルグラムの平面性も線形時間で判定できることが分かる．

後に Lozano, Pinter, Rokhlenko, Valiente, and Ziv-Ukelson [12] も同じ問題に対する多項式時間アルゴリズムを独立に与えている．彼らのアルゴリズムの計算量は $O(n^2)$ である．

4 交差数最小化の難しさ

前節で見たように，二分タングルグラムが辺交差なく描けるかどうかの判定は線形時間で可能である．それでは辺交差の最小化はどれ程難しいのだろうか．実はこの問題が NP 困難となることを Fernau, Kaufmann, and Poths [6] は証明している．帰着には最大カット問題を用いている．

そのため，問題が簡単に解けるような部分クラスを考えたくなる．しかし，Buchin, Buchin, Byrka, Nöllenburg, Okamoto, Silveira, and Wolff [2] は 2 つの二分木が完全二分木であっても BTL が NP 困難であることを示した．帰着には Max2SAT を用いている．完全二分木は最も簡単な特殊ケースに思えるため，それでも BTL が NP 困難になることは興味深い．

5 近似アルゴリズム

BTL が NP 困難であると分かったので，多項式時間近似アルゴリズムに対する興味が湧き上がる．近似アルゴリズムの近似の良さは次で定義される近似比で測られる．BTL に対して (より一般的に最小化問題に対して) アルゴリズム A が α 近似であるとは，任意のインスタンスに対して A の出力する許容解の目的関数値が最適値の α 倍以下となることである．ただし，最適値は 1 以上であることを仮定する (実際，BTL は最適値が 0 であるかどうか線形時間で判定できるので，この仮定は妥当である)．この $\alpha \geq 1$ はアルゴリズム A の近似比と呼ばれる．目標は $\alpha \geq 1$ がより小さいアルゴリズムを設計することである．

しかし，Buchin, Buchin, Byrka, Nöllenburg, Okamoto, Silveira, and Wolff [2] は，一意ゲーム予想 (unique games conjecture) が成立するとき，任意の定数 α に対して BTL の α 近似解を見つけることが NP 困難であることを証明した．一意ゲーム予想とは Khot [11] による予想で，その成立は P$\neq$NP ほど強く信じられていないが (しかし，P$\neq$NP が成り立てば一意ゲーム予想も

成り立つという含意は成り立つ)，様々な組合せ最適化問題に対する近似比のタイトな下界をもたらすため成立を信じる研究者も多い．

そこで，次の 2 つの問題設定を考える．1 つは目的関数を変えることである．すなわち，辺交差を最小化するのではなく，交差しない葉の対の数を最大化するのである．実を言うと，このバージョンは定数近似比の多項式時間アルゴリズムを持つ．詳細は省くが，Buchin, Buchin, Byrka, Nöllenburg, Okamoto, Silveira, and Wolff [2] はこの問題を最大カット問題 (の変種) として定式化し，それに対する Goemans and Williamson [8] のアルゴリズムを用いることで，近似比が 0.878 の多項式時間アルゴリズムを提案している (これは最大化問題であり，上で定義している近似比と異なる定義が適用されているので注意が必要である)．これは与えられたタングルグラムが二分木でなくても同じように適用できる．

もう 1 つの問題設定は，与えられる二分木がともに完全二分木である場合である．先に述べたように，このバージョンも NP 困難である．しかし，一般の BTL と違い，完全二分木に対する BTL に対して Buchin, Buchin, Byrka, Nöllenburg, Okamoto, Silveira, and Wolff [2] は近似比が 2 の多項式時間アルゴリズムを設計した．以下はその概要である．BTL のインスタンス $\langle S, T\rangle$ が与えられたとする．S の根の子を根とする 2 つの部分木を S_1, S_2 として，T の根の子を根とする 2 つの部分木を T_1, T_2 とする．これら 4 つの部分木を配置する方法は 4 つある．(1) S_1 を S_2 の左に置き，T_1 を T_2 の左に置く．(2) S_1 を S_2 の左に置き，T_1 を T_2 の右に置く．(3) S_1 を S_2 の右に置き，T_1 を T_2 の左に置く．(4) S_1 を S_2 の右に置き，T_1 を T_2 の右に置く．そして，この 4 つの場合それぞれに対して向かい合う部分木どうしから成る 2 つのインスタンスを再帰的に解く．その子問題から得られる交差数と配置の選択自身から得られる交差数の和が最小なものを出力とする．これがアルゴリズムの概要である．

彼らのアルゴリズムは与えられた木が完全 d 分木の場合にも適用でき，その場合の近似比は $1 + \binom{d}{2}$ となる．また，アルゴリズムは少しの変更で完全とは限らない二分木にも適用できる．しかし，その場合近似比の保証は得られない．これについては Nöllenburg, Völker, Wolff, and Holten [14] が実験的解析を行っている．

6 固定パラメータ・アルゴリズム

NP 困難な問題に対するアプローチとして，近似アルゴリズムとは別の方向性に固定パラメータ・アルゴリズムというものがある．これは問題を困難にする内在パラメータを特定し，それに対する計算量の依存性を小さくするようなアルゴリズムを設計していくアプローチである．特に，問題を厳密に解くことを主眼にすることが多い．

パラメータ k を持ち，インスタンスのサイズが n である問題に対する固定パラメータ・アルゴリズム (fixed-parameter algorithm) とは，その計算量がある関数 f と多項式 p を用いて $O(f(k)p(n))$ と書けるようなアルゴリズムである．ここで，f は多項式でなくてもよいことに注意する．

ここから考える BTL に対する固定パラメータ・アルゴリズムにおいてパラメータ k は最適値である．BTL に対して，Fernau, Kaufmann, and Poths [6] は $O(c^k p(n))$ 時間の固定パラメータ・アルゴリズムを与えた．ただし，p はある多項式で c はおよそ 2^{10} となる定数である．

Buchin, Buchin, Byrka, Nöllenburg, Okamoto, Silveira, and Wolff [2] は完全二分木に対する BTL に限るが，Fernau, Kaufmann, and Poths [6] のものに比べて各段にシンプルな固定パラメータ・アルゴリズムを与えた．その計算量は $O(4^k n^2)$ である．基本的な戦略は有界探索木 (bounded search tree) によるが，探索木においてパラメータが常に減少するわけではないところに特徴がある．

7 実験

Nöllenburg, Völker, Wolff, and Holten [14] はこれまでに提案されたアルゴリズムに加えて，新たなアルゴリズムを提案し，実験的評価を行った．彼らが考察対象にしたアルゴリズムは以下のものである．

(1) 上で概略を説明した Buchin, Buchin, Byrka, Nöllenburg, Okamoto, Silveira, and Wolff [2] による完全二分タングルグラムに対する 2 近似アルゴリズムを一般の二分タングルグラムに対しても動くように拡張したもの．簡単なヒューリスティクスも加えている．

(2) Holten and van Wijk [10] によるアルゴリズム．これは Sugiyama, Tagawa, and Toda [16] による 1SCM に対するアルゴリズムをサブルーティンとして用い，一方の木を固定した後，もう一方の木のレイアウトを定めるというステップを木の役割を交代させながら繰り返すというものである．

(3) Dwyer and Schreiber [4] によるアルゴリズム．これは (2) のアルゴリズムに似ているが，Sugiyama, Tagawa, and Toda [16] のアルゴリズムによってもう一方の木のレイアウトを定めずに，上で説明した片面バージョンに対する厳密アルゴリズムを用いる．

(4) 整数計画法として定式化．それを CPLEX によって解く．これは厳密解法である．

(5) 分枝限定法．(1) のアルゴリズムにある分割の考えを発展させることで単純な分枝限定アルゴリズムを構成している．これも厳密解法である．

(6) 分枝限定法による初期解．(5) の分枝限定法によって発見された最初の暫定解をそのまま出力する発見的解法．

彼らは，ランダムに構成した二分木に対する実験と実データに対する実験を行っている．問題規模は葉の数が 600 近くのものまでである．その中ですべての場合に対して (6) が極めて最適に近い解を出力している．(4) と (5) の厳密アルゴリズムはすべてのインスタンスを 1 分以内に解けるわけではなく，特に (4) は交差数の多いランダムな大規模インスタンス，(5) は交差数の少ないランダムな大規模インスタンスに弱い．実行時間と解の近似精度の観点から (6) のヒューリスティクスが実用上優れていると彼らの論文は結論づけている．

8 結語

情報可視化において二分タングルグラムを描くことの重要さは増加していくと思われる．その一方で，二分タングルグラム描画に対するアルゴリズム的研究が大きく進展しているとは思えない．より詳細な議論とその成果を生かした情報可視化ツールの作成が大きな研究問題として残されている．

謝辞 筆者のタングルグラムに対する興味と研究成果は Kevin Buchin, Maike Buchin, Jaroslaw Byrka, Martin Nöllenburg, Rodrigo I. Silveira, Alexander Wolff との共同研究 [2] によるものである．彼らに深く感謝する．

参考文献

[1] P. Bertolazzi, G. Di Battista, C. Mannino, and R. Tamassia. Optimal upward planarity testing of single-source digraphs. *SIAM J. Comput.*, 27(1):132–169, 1998.

[2] K. Buchin, M. Buchin, J. Byrka, M. Nöllenburg, Y. Okamoto, R. I. Silveira, and A. Wolff. Drawing (complete) binary tanglegrams: Hardness, approximation, fixed-parameter tractability. In I. Tollis and M. Patrignani, editors, *Proc. 16th Internat. Sympos. Graph Drawing (GD'08)*, volume 5417 of *Lecture Notes Comput. Sci.*, pages 324–335. Springer-Verlag, 2009. Preprint available at http://arxiv.org/abs/0806.0920.

[3] V. Dujmović, H. Fernau, and M. Kaufmann. Fixed parameter algorithms for one-sided crossing minimization revisited. In G. Liotta, editor, *Proc. 11th Internat. Sympos. Graph Drawing (GD'03)*, volume 2912 of *Lecture Notes Comput. Sci.*, pages 332–344. Springer-Verlag, 2004.

[4] T. Dwyer and F. Schreiber. Optimal leaf ordering for two and a half dimensional phylogenetic tree visualization. In N. Churcher and C. Churcher, editors, *Proc. Australasian Sympos. Inform. Visual. (InVis.au'04)*, volume 35 of *CRPIT*, pages 109–115. Australian Comput. Soc., 2004.

[5] P. Eades and N. Wormald. Edge crossings in drawings of bipartite graphs. *Algorithmica*, 10:379–403, 1994.

[6] H. Fernau, M. Kaufmann, and M. Poths. Comparing trees via crossing minimization. In R. Ramanujam and S. Sen, editors, *Proc. 25th Intern. Conf. Found. Softw. Techn. Theoret. Comput. Sci. (FSTTCS'05)*, volume 3821 of *Lecture Notes Comput. Sci.*, pages 457–469, 2005.

[7] M. R. Garey and D. S. Johnson. Crossing number is NP-complete. *SIAM Journal on Algebraic and Discrete Methods*, 4:312–316, 1983.

[8] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM*, 42(6):1115–1145, 1995.

[9] M. S. Hafner, P. D. Sudman, F. X. Villablanca, T. A. Spradling, J. W. Demastes, and S. A. Nadler. Disparate rates of molecular evolution in cospeciating hosts and parasites. *Science*, 265:1087–1090, 1994.

[10] D. Holten and J. J. van Wijk. Visual comparison of hierarchically organized data. In *Proc. 10th Eurographics/IEEE-VGTC Sympos. Visualization (EuroVis'08)*, pages 759–766, 2008.

[11] S. Khot. On the power of unique 2-prover 1-round games. In *Proc. 34th Annu. ACM Sympos. Theory Comput. (STOC'02)*, pages 767–775, 2002.

[12] A. Lozano, R. Y. Pinter, O. Rokhlenko, G. Valiente, and M. Ziv-Ukelson. Seeded tree alignment and planar tanglegram layout. In R. Giancarlo and S. Hannenhalli, editors, *Proc. 7th Internat. Workshop Algorithms Bioinformatics (WABI'07)*, volume 4645 of *Lecture Notes Comput. Sci.*, pages 98–110. Springer-Verlag, 2007.

[13] H. Nagamochi. An improved bound on the one-sided minimum crossing number in two-layered drawings. *Discrete Comput. Geom.*, 33(4):565–591, 2005.

[14] M. Nöllenburg, M. Völker, A. Wolff, and D. Holten. Drawing binary tanglegrams: An experimental evaluation. In *Proc. 10th Workshop on Algorithm Engineering and Experiments (ALENEX'09)*, pages 106–119, 2009. Preprint available at http://arxiv.org/abs/0806.0928.

[15] R. D. M. Page, editor. *Tangled Trees: Phylogeny, Cospeciation, and Coevolution.* University of Chicago Press, 2002.

[16] K. Sugiyama, S. Tagawa, and M. Toda. Methods for visual understanding of hierarchical system structures. *IEEE Transactions on Systems, Man, and Cybernetics*, 11(2):109–125, 1981.

[17] A. Yamaguchi and A. Sugimoto. An approximation algorithm for the two-layered graph drawing problem. In T. Asano, H. Imai, D. T. Lee, S. ichi Nakano, and T. Tokuyama, editors, *Proc. 5th Annu. Internat. Comput. Combinatorics Conf. (COCOON'99)*, volume 1627, pages 81–91, 1999.